Foundations of Human-Agent Collaboration: Situation-Relevant Information Sharing

Tim Miller*, **Adrian Pearce***, **Liz Sonenberg***,
Frank Dignum**, **Paolo Felli*** and **Christian Muise***
*Department of Computing and Information Systems, University of Melbourne
**Department of Information and Computing Sciences, Universiteit Utrecht
{tmiller,adrianrp,l.sonenberg,paolo.felli,christian.muise}@unimelb.edu.au, F.P.M.Dignum@uu.nl

Introduction

Empirical studies with humans and agents demonstrate that the nature and forms of information required by the human differ depending on the design of the relationship between the participants — a relationship that is sometimes characterised using the concept of levels of autonomy (Parasuraman, Sheridan, and Wickens 2000), though the usefulness of that characterisation has recently been questioned (DSB 2012). Therefore, understanding how people work with automation and how to design automated systems to better support people, is a field long studied, but of growing importance (Birmingham and Taylor 1954; Hancock et al. 2013). Our current work seeks to contribute to the design of representations and algorithms that can be deployed in such contexts.

Our goal is not to make agents more autonomous, but to make them more capable of being *interdependent* (Johnson et al. 2014), where interdependent informally means that the choice and outcome of an agent's action is dependent on what another agent does, and vice-versa. It is argued that agents that could act interdependently with humans would enable a more natural interaction between humans and agents.

Agents are more capable of being interdependent if each has awareness of what is happening with the other(s). In the context of human-agent collaboration, this leads to a requirement that agents and humans are capable of reasoning about others and the relevant context, known as *social reality* (Dignum, Prada, and Hofstede 2014), including their awareness of the situation, and their possible behaviours. Therefore, an intelligent agent must have a model of the human, and a model of the human's model of the agent itself and other participants in an activity, and possibly such nested models down several levels. This can enable agents to combine temporal and epistemic projection to predict future actions of others (Pearce, Sonenberg, and Nixon 2011).

Our overarching focus is on the challenges of *interpredictability* and establishing *common ground*, marked as fundamental in recent commentaries on human-automation research (Hancock et al. 2013; Johnson et al. 2014). Interpredictability refers to the knowledge, coordination devices (e.g. explicit and implicit communication), working agreements and forms of feedback needed to achieve collaboration and avoid surprises. Common ground refers to the mutually shared belief, acceptances, and assumptions that support interdependent actions in the context of a joint activity (Pfau et al. 2014). Our approach seeks to combine logic-based approaches with recent advances in planning.

Modelling Common Ground

It has been argued that establishing common ground between humans and artificial agents can improve their collaborative efforts, e.g., (Klein et al. 2004). However, research in artificial intelligence has not provided a precise definition of common ground that is suitable for computation. Moreover, informal conceptions of the notion vary substantially.

In recent work (Pfau et al. 2014), some of the authors defined four logics of common ground. Starting with a straightforward definition of common ground as *common belief*, the definitions progress to be more aligned with how human teams establish common ground, modelling existing definitions in philosophy and social psychology. Specifically, these models use the notion of *acceptance*: a mental attitude similar to belief, but which allows us to accept things as true without actually believing them, for the purpose of completing a task. In addition, we presented a new definition of common ground, called *salient common ground*, which models how people build up common ground in a particular activity using three sources: (1) *activity-specific common ground*: the common ground specific to the current activity; (2) *personal common ground*: the common ground held between the group members built from previous personal experience with each other; and (3) *communal common ground*: the common ground held between the group members based on some mutual membership of a community (e.g. nationality, religious groups, groups who have trained together).

Formally modelling existing accounts of common ground allowed us to show how some existing accounts are equivalent, despite seeming different, and allowed us to eliminate the ambiguities inherent in such informal definitions.

Epistemic Planning

Reasoning about action and change has typically focused on how the world changes as a result of an agent's behaviour.

An aspect of our research aims to extend this reasoning to model how the *belief* of an agent changes as a result of actions in the world, including how their nested beliefs change. That is, beliefs about the beliefs of other agents, including what an agent a believes another agent b believes about the world and about a's beliefs. Modelling the knowledge or belief of interacting agents is crucial for many scenarios that contain a dynamic environment, and we are primarily interested in how to plan for achieving goals that involve arriving at a particular mental state for a group of agents.

Previous work has focused on how epistemic planning can be addressed from a theoretical standpoint by appealing to dynamic epistemic logic (Bolander and Andersen 2011). In our work, we take the complementary approach of modelling belief in a manner amenable to automated planning techniques. In particular, we consider a syntactic restriction of an agent's mental model, represented using traditional non-deterministic planning techniques (Muise, McIlraith, and Beck 2012).

Social Reality

Recently, one of the authors proposed a high-level cognitive framework (Dignum, Prada, and Hofstede 2014) for social agents – that is, those agents that have "sociality" as part of the core reasoning. One aspect of this framework is the concept of *social reality*, which refers to the notion that people have a model of social behaviour that they use to reason about others. For example, when interacting with a bank manager, we use our model of bank managers to reason that the person will know certain things, and to reason about what they will do if we give them a certain request.

An agent can reason about others by having models of those others, however, having detailed models of their beliefs, acceptances, desires, and capabilities is not required in many cases, and in fact, may be detrimental due to the computational cost. When interacting with a bank teller, an agent can take a general model of a bank manager and use this to approximate the manager's core values and behaviours.

Current work focuses on basic computational models of social reality. Specifically, we are defining a model that permits agents to use *stereotypical reasoning* about another agent using simple social rules, or to use *empathetic reasoning* about another agent; that is, to cast itself into the mind of the other agent and reason about what it would do. Some aspects of this type of reasoning have been explored (Pfau, Kashima, and Sonenberg 2014). There are potential links between the logic-based mechanisms explored in that work, and prior planning-oriented work on composition of goal-based processes where human actors in the environment are abstracted as services and wrapped by a semantic description allowing them to be involved in orchestrations, and to collaborate with devices, to reach certain goals (De Giacomo et al. 2012). As well as saving computational effort, stereotypical reasoning allows the agent to reason about another agent playing a role, even if it has no model of the agent's current mental states or capabilities. Our approach allows for the fusion of both; e.g., having a model of the agent's current beliefs, and employing a stereotype rule that accesses those beliefs to reason about the agent.

Current and Planned Work

In scenarios where others' actions are predictable, complex mechanisms for online adaptation are not needed. Our work seeks to provide fundamental mechanisms for dynamic environments where agents must adapt 'autonomously.' Our current focus is on computationally tractable models of nested belief/knowledge for multi-agent settings, with particular emphasis on social reality. We are using logic-based and planning-oriented mechanisms. Future work will look at building tools that enable experimentation in scenarios with human players.

Acknowledgements: This research is partially funded by Australian Research Council Discovery Grant DP130102825.

References

Birmingham, H., and Taylor, F. V. 1954. A design philosophy for man-machine control systems. *Proceedings of the IRE* 42(12):1748–1758.

Bolander, T., and Andersen, M. 2011. Epistemic planning for single and multi-agent systems. *Journal of Applied Non-Classical Logics* 21(1):9–34.

De Giacomo, G.; Ciccio, C. D.; Felli, P.; Hu, Y.; and M, M. 2012. Goal-based composition of stateful services for smart homes. In *On the Move to Meaningful Internet Systems: OTM 2012*, volume 7565 of *LNCS*, 194–211. Springer.

Dignum, F.; Prada, R.; and Hofstede, G. 2014. From autistic to social agents. In *AAMAS Proceedings*, 1161–1164.

DSB. 2012. The Role of Autonomy in DoD Systems. 125pp, http://www.acq.osd.mil/dsb/reports/AutonomyReport.pdf.

Hancock, P.; Jagacinski, R.; Parasuraman, R.; Wickens, C.; Wilson, G.; and Kaber, D. 2013. Human-automation interaction research: Past, present, and future. *Ergonomics in Design: The Quarterly of Human Factors Applications* 21(2):9–14.

Johnson, M.; Bradshaw, J. M.; Feltovich, P. J.; Jonker, C. M.; van Riemsdijk, M. B.; and Sierhuis, M. 2014. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction* 3(1):43–69.

Klein, G.; Hoffman, R. R.; Feltovich, P. J.; Woods, D. D.; and Bradshaw, J. M. 2004. Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems* 19(6):91–95.

Muise, C.; McIlraith, S.; and Beck, J. 2012. Improved Non-deterministic Planning by Exploiting State Relevance. In *ICAPS*.

Parasuraman, R.; Sheridan, T. B.; and Wickens, C. D. 2000. A model for types and levels of human interaction with automation. *Trans. Sys. Man Cyber. Part A* 30(3):286–297.

Pearce, A.; Sonenberg, L.; and Nixon, P. 2011. Toward resilient human-robot interaction through situation projection for effective joint action. In *Robot-Human Teamwork in Dynamic Adverse Environment: AAAI Fall Symp.*, 44–48.

Pfau, J.; Miller, T.; Sonenberg, L.; and Kashima, Y. 2014. Logics of common ground. (Under review), http://people.eng.unimelb.edu.au/tmiller/pubs/logics-of-common-ground.pdf,.

Pfau, J.; Kashima, Y.; and Sonenberg, L. 2014. Towards agent-based models of cultural dynamics: A case of stereotypes. In *Perspectives on Culture and Agent-based Simulations*. 129–147.